논문 2011-48SP-5-5

군집 주제의 유의어와 유사도를 이용한 문서군집 향상 방법

(Enhancing Document Clustering Method using Synonym of Cluster Topic and Similarity)

박 선*, 김 경 준**, 이 진 석***, 이 성 로****

(Sun Park, Kyung-Jun Kim, Jin-Seok Lee, and Seong Ro Lee)

요 약

본 논문은 군집 주제의 유의어와 유사도를 이용하여 문서군집의 성능을 향상시키는 방법을 제안한다. 제안된 방법은 비음수 행렬분해의 의미특징을 이용하여 군집 주제(topic)의 용어들을 선택함으로서 문서 군집 집합의 내부구조를 잘 표현할 수 있으며, 군집 주제의 용어들에 워드넷의 유의어를 사용하여서 확장함으로써 문서를 용어집합(BOW, bag-of-words)으로 표현하는 문제를 해결할 수 있다. 또한 확장된 군집 주제의 용어와 문서집합에 코사인 유사도를 이용하여서 군집의 주제에 적합한 문서를 잘 군집하여서 성능을 높일 수 있다. 실험결과 제안방법을 적용한 문서군집방법이 다른 문서군집 방법에 비하여 좋은 성능을 보인다.

Abstract

This paper proposes a new enhancing document clustering method using a synonym of cluster topic and the similarity. The proposed method can well represent the inherent structure of document cluster set by means of selecting terms of cluster topic based on the semantic features by NMF. It can solve the problem of "bags of words" by using of expanding the terms of cluster topics which uses the synonyms of WordNet. Also, it can improve the quality of document clustering which uses the cosine similarity between the expanded cluster topic terms and document set to well cluster document with respect to the appropriation cluster. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Keywords : 문서군집(document clustering), 비음수행렬분해(NMF, non-negative matrix factorization), 의미 특징(semantic features), 유의어(synonym), 코사인 유사도(cosine similarity).

Ⅰ. 서 론

문서군집은 정보검색, 문서요약, 자동문서 조직, 주제 추출, 정보 필터링 등에 효율적인 기반 기술로서 많이 사용된다. 특히 트위터, 패이스북, 블로그, 온라인 뉴스 등의 많은 문자자료들이 급속도로 늘어남에 따라서 더욱 많은 관심이 집중되고 있다[1~2].

일반적인 군집방법은 분할기반 방법, 계층적 기반 방법, 밀도기반 방법, 격자 기반 방법으로 분류 할 수 있다. 이들 중 문서 군집에서 자주 사용되는 방법으로는 분할기반 방법과 계층적 기반 방법이 있다. 분할기반

* 정회원-교신저자, 목포대학교 정보산업연구소
(Institute of Information Science and Engineering Research, Mokpo National University)
** 정회원, 한국과학기술원 전산학과
(Departmnet of Computer Science, KAIST)
*** 정회원, 정보통신산업진흥원
(NIPA)
**** 정회원, 목포대학교 정보전자공학과
(Department of Information and Electronics)
※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임(2010-0028295)
접수일자: 2011년3월16일, 수정완료일: 2011년6월9일

방법은 k개의 초기 군집을 생성한 다음, 군집의 성능을 향상시키기 위하여 반복적으로 군집의 객체들을 제배치하는 방법이다. 계층적 기반 방법은 주어진 자료 집합을 계층적 트리형태로 군집하는 방법이다. 그러나 이러한 군집방법들은 대부분 거리 기반의 목적 함수를 사용하기 때문에 고차원의 객체들을 군집해야하는 문서 군집에는 비효율적이다[2~4].

문서군집 방법을 정의하면 군집 알고리즘을 사용하여서 유사한 특성의 문서들을 집합으로 묶는 기술이다[3~4]. 일반적인 문서군집 방법들에서는 문서를 용어집합(BOW, bag-of-words)으로 표현하는 방법을 주로 사용한다. 그러나 이러한 문서를 용어의 집합으로 표현하는 방법은 문서 집합에 포함된 용어들의 의미적 관계를 전혀 고려하지 않고, 단지 용어들이 문서에 출현된 빈도만을 이용한다. 이 때문에 용어 빈도 집합을 사용한 문서군집 방법은 문서 집합 내에 포함된 문서의 특성이 군집의 결과에 많은 영향을 미친다. 즉, 문서들의 분포나 문서집합의 내부구조, 사용자가 요구하는 군집형태 등에 따라서 군집의 결과가 달라진다. 또한 문서 집합을 군집 할 때에 문서간의 거리를 측정하여서 군집하는 거리기반의 목적함수를 사용함으로써 두 문서간의 실제 거리를 잘 반영할 수 없다[2]. 이러한 문제를 해결하기 위해서 요즘은 문서군집에서는 온톨로지(ontology, 공유된 개념화)나 의미특징(semantic feature)을 이용한 방법을 많이 사용하고 있다.

온톨로지에 기반한 문서군집 방법은 워드넷이나 위키피디아 등의 외부 지식으로부터 용어 온톨로지를 구축하여서 문서군집의 성능을 향상시킨다. 그러나 문서집합에서 사용되는 용어에 대해서 포괄적인 개념을 찾아 온톨로지를 구축하는 것이 어렵다. 또한 온톨로지 구축비용이 많이 들고, 온톨로지를 구축하더라도 정확한 범위를 적용대상에 적용하는 것도 힘들다. 또한 이러한 특성 때문에 때로는 정보손실 문제가 발생한다[2]. 최근의 온톨로지를 이용한 문서군집에 대한 연구로는 다음과 같다. Hu외 저자들은 문서군집을 위하여 위키피디아의 외부 지식을 이용하여 온톨로지를 구축하였다[2]. Trappey외 저자들은 특허문서를 군집하기 위하여 퍼지에 기반한 온톨로지 방법을 제안하였다[5].

의미특징에 기반한 문서군집 방법은 문서집합 내부의 특성을 나타내는 의미특징을 이용한 방법으로서 쉽게 군집의 특성을 나타내는 주제들을 추출할 수 있다. 또한 명확한 의미의 군집주제를 나타내는 의미특징을 이용하여 좋은 군집결과를 얻을 수 있다. 그러나 문서집합의 구성 문서들이 유사한 특성을 같거나, 극단적으로 다른 특성을 갖고 있으면 추출된 의미특징들의 문서집합의 내부 구조를 충분하게 반영할 수 없으므로 좋은 군집 결과를 얻기 힘들다[6]. 의미특징에 기반한 문서군집의 최근 연구는 다음과 같다. Li 이외 저자들은 문서군집과 관련된 군집의 하위 공간구조의 특징을 이용한 ASI(Adaptive Subspace Iteration) 알고리즘을 제안하였다[7]. Wang과 Zhang은 문서군집을 위하여 지역 레이블과 전역 레이블의 특징을 이용한 CLGR(Clustering with Local and Global Regularization) 알고리즘을 제안하였다[8]. Xu이외 저자들은 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)의 의미특징을 이용하여 문서를 군집하는 방법을 제안하였다[9]. 본 논문의 저자들은 이전에 문서군집을 위한 세 가지 방법을 제안하였다. 제안방법으로는 의미특징과 군집의 응집도를 이용한 방법[10~11], 의미특징과 퍼지관계를 이용한 방법[12], 마지막으로 주성분 분석과 퍼지연관을 이용한 방법[13]이 있다. 이들 방법은 의미특징에 기반을 두고 있기 때문에 구성 문서의 특성이 극단적으로 유사하거나 다르면 군집의 성능이 좋지 않을 수 있는 문제를 가지고 있다.

본 논문에서는 의미특징 방법의 제한 사항을 극복하기 위하여서 군집 주제의 유의어와 코사인 유사도를 이용한 문서군집의 성능 향상 방법을 제안한다. 제안 방법은 다음과 같다. 첫 번째 단계로 비음수 행렬의 의미특징을 이용하여서 군집의 주제를 나타낼 수 있는 중요도가 높은 용어들을 추출한다. 이렇게 추출된 용어들은 군집의 내부 특성을 잘 반영할 수 있는 군집의 주제를 요약된 형태로 잘 표현할 수 있다. 두 번째 단계에서는 추출된 용어들을 워드넷을 이용하여서 유의어로 확장한다. 확장된 군집 주제의 용어들은 의미특징이 원본 문서집합의 문서구성에 제한받는 문제를 극복할 수 있다. 마지막으로 확장된 군집 주제의 용어들과 원본 문서들 간에 코사인 유사도를 이용하여서 문서를 군집한다. 군집 주제를 잘 반영 할 수 있는 문서들을 코사인 유사도를 이용하여서 군집함으로써 군집의 성능을 향상 시킬 수 있다.

본 논문의 구성은 다음과 같다. 제Ⅱ장은 관련연구로 비음수행렬분해를, 제Ⅲ장은 제안한 문서군집 방법을, 제Ⅳ장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제Ⅴ장에서는 결론을 맺는다.

Ⅱ. 비음수행렬분해

이번 장에서는 비음수행렬분해의 개념과 알고리즘에 대하여 알아보고, 다음 장에서 비음수행렬분해를 이용하여서 군집 주제의 중요 용어를 추출하는 제안방법에 대하여 알아본다.

비음수행렬분해는 대량의 객체정보로부터 비음수로 된 부분정보를 추출하고, 이들의 선형 조합으로 객체를 표현할 수 있도록 하는 방법이다. 비음수행렬분해 알고리즘은 비음수 자료로 구성된 원본 자료를 두 개의 비음수로된 행렬로 분해한다[6]. 비음수 행렬 분해 알고리즘은 식(1)의 목표함수 J 가 0에 가깝게 수렴 할 때까지 식(2)과 식(3)을 이용하여 행렬 W와 H의 값을 동시에 갱신한다.

$$J = \| A - WH \|^2 \tag{1}$$

식(1)의 목적은 행렬 A를 비음수 $m \times r$ 행렬 W와 비음수 $r \times n$ 행렬 H로 분해하는 것이다. 여기서, A는 m개의 용어와 n개의 문장으로 이루어진 $m \times n$ 행렬이고, r은 의미특징행렬의 크기를 결정할 수 있는 의미특징의 개수이다. 또한 두 개의 비음수 의미 특징 행렬을 구별하기 위하여서, 비음수행렬분해 알고리즘을 제안한 Lee와 Seung은 두 행렬 W와 H를 의미특징 행렬 W와 의미변수 행렬 H로 각각 이름을 정의하였다[6].

$$H_{rj} \leftarrow H_{rj} \frac{(W^T V)_{rj}}{(W^T W H)_{rj}} \tag{2}$$

$$W_{ir} \leftarrow W_{ir} \frac{(V H^T)_{ir}}{(W H H^T)_{ir}} \tag{3}$$

다음 그림 1은 비음수행렬분해 알고리즘을 이용하여 비음수 행렬 A를 두개의 비음수 행렬 W와 H로 분해하는 예이다. 그림 1의 행렬 A는 Matlab 7.8의 *nnmf*()함수를 이용하여서 행렬분해를 하였다. 여기서 비음수행렬분해로 분해된 행렬 W와 H는 PCA(principal components analysis)나 VQ(vector quantization)과 같은 다른 행렬분해 알고리즘의 결과와 비교하여서 0 값을 많이 포함한 희소행렬을 이루는 것을 알 수 있다. 이렇게 의미특징이 희소행렬을 구성하면 문서집합의 주제를 더 적은 수의 의미특징으로 표현할 수 있다. 즉, 몇몇 중요한 용어들만으로 군집에서 나타내고자 하는 주제를 표현할 수 있다.

$$\underset{A}{\begin{bmatrix}0&0&1&2\\0&1&0&0\\0&0&0&1\\1&0&0&3\\0&0&2&1\end{bmatrix}} \approx \underset{W}{\begin{bmatrix}1.395&1.188&0\\0&0&1.000\\0.949&0.038&0\\3.169&0&0\\0&2.262&0\end{bmatrix}} \times \underset{H}{\begin{bmatrix}0.254&0&0&0.967\\0&0&0.883&0.469\\0&1.000&0&0\end{bmatrix}}$$

그림 1. 비음수행렬분해 알고리즘의 결과
Fig. 1. The result of NMF algorithm.

$$\underset{A_{*4}}{\begin{bmatrix}2\\0\\1\\3\\1\end{bmatrix}} \approx 0.967 \times \underset{H_{14}}{\underset{W_{*1}}{\begin{bmatrix}1.395\\0\\0.949\\3.169\\0\end{bmatrix}}} + 0.469 \times \underset{H_{24}}{\underset{W_{*2}}{\begin{bmatrix}1.188\\0\\0.038\\0\\2.262\end{bmatrix}}} + 0 \times \underset{H_{34}}{\underset{W_{*3}}{\begin{bmatrix}0\\1.000\\0\\0\\0\end{bmatrix}}}$$

그림 2. 의미특징 행렬을 이용한 문서의 표현
Fig. 2. The representing of document by semantic feature matrix.

본 논문에서 행렬 X의 j번째 열벡터는 X_{*j}로, i번째 행벡터는 X_{i*}로, i번째 행과 j번째 열의 원소는 X_{ij}표시한다. 행렬 A의 j번째 열벡터 A_{*j}는 행렬 W의 l번째 열벡터 W_{*l}와 행렬 H의 요소 H_{kj}가 선형조합을 이루며 식(4)과 같다[6].

$$A_{*j} = \sum_{l=1}^{r} H_{kj} W_{*l} \tag{4}$$

다음 그림 2는 식(4)와 같은 형태로 비음수행렬분해된 의미특징을 이용하여 문서를 표현하는 예이다. 여기서는 그림 1의 4번째 열벡터인 A_{*4}를 하나의 문서로 가정할 때에 그림 2와 같이 문서 A_{*4}를 의미특징벡터와 의미변수의 선형조합으로 나타낼 수 있다.

Ⅲ. 제안 문서군집 방법

본 논문에서 제안한 문서군집 과정은 다음 그림 3과

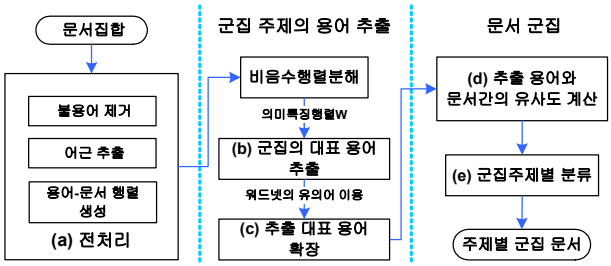

그림 3. 의미 특징과 유사도를 이용한 문서군집
Fig. 3. Document clustering by semantic feature and similarity.

같이 전처리, 군집 주제의 용어 추출, 문서군집으로 구성된다. 전처리단계에서는 문서집합을 전처리하여서 용어-문서 빈도행렬을 구성한다. 군집 주제의 용어 추출 단계에서는 비음수행렬분해를 이용하여 군집의 주제를 요약하여서 설명할 수 있는 중요 용어들을 추출하고, 워드넷을 이용하여 추출된 용어들을 확장한다. 문서군집단계에서는 추출된 군집주제의 용어 집합과 문서들 간의 유사도를 계산하여 문서를 군집한다.

3.1 전처리

일반적으로 문서군집의 성능평가에서 사용되는 표준 평가 자료[7~9]들은 대부분 영문 문서로 구성되어있다. 이 때문에 본 논문의 전처리는 영문문서를 처리하는 방법을 기준으로 설명한다. 만약 한글 문서에 본 논문에서 제안방법을 적용하려면, 전처리 단계만 한글 형태소 분석 도구[14]를 사용하여 용어만 추출한 다음에 용어-문서 빈도행렬을 구성하면 된다.

그림 3(a)의 전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[3, 5, 15]. 불용어 제거는 Rijsbergen의 불용어 목록[15]을 이용하여서 목록에서 정의하고 있는 무의미한 용어들을 제거한다. 어근추출은 Porter의 어근추출 알고리즘[15]을 이용하여서 영어의 파생어들을 가장 중심이 되는 용어인 어근으로 변환한다. 용어-문서 빈도 행렬의 용어빈도 벡터 생성에 사용되는 벡터 $T_i = [\ t_{1i},\ t_{2i},\ \cdots,\ t_{ni}\ ]^{\mathrm{T}}$는 i번째 문장의 용어빈도이다. 여기서 요소 t_{ij}는 j번째 문서에서 출현한 i번째 용어의 빈도이다[3, 5, 15].

3.2 군집주제의 용어 추출

이 장에서는 비음수행렬분해를 이용하여 그림 3(b)와 같이 군집 주제를 나타내는 군집의 대표 용어들을 추출하고, 그림 3(c)와 같이 추출된 군집의 대표 용어를 워드넷의 유의어를 이용하여 확장한다.

군집주제의 용어를 추출할 때 비음수 행렬분해를 사용하는 이유는, 비음수행렬분해에 의해 생성되는 의미특징들은 원본 문서집합의 내부구조를 잘 반영하여 나타낼 수 있다. 즉, 일반적으로 문서집합은 다양한 주제를 갖는 문서들로 구성되어 있고, 각각의 주제를 포함하는 문서들을 모아서 군집을 구성할 수 있다. 이 때문에 비음수행렬분해 된 의미특징들은 이러한 문서들이 포함하고 있는 중요한 주제들을 쉽게 의미 있는 특징들로 그룹화하여서 나타낼 수 있다.

그러나 의미특징들은 문서집합의 내부의 구조특성만을 이용하기 때문에 실제로 문서들이 같은 주제를 포함하고 있으면서 다른 형태로 표현하는 경우 잘 구분할 수 없는 문제를 가지고 있다. 이러한 문제를 해결하기 위하여서 본 논문에서는 영어 어휘사전이 워드넷을 이용하여서 군집주제 용어를 유의어로 확장하여 사용한다. 용어를 유의어로 확장하면, 같은 주제를 다르게 표현하더라도 유사한 용어로 나타내기 때문에 쉽게 이를 구분할 수 있다.

가. 군집의 대표 용어 추출

그림 3(b)와 같이 군집의 주제를 잘 요약하여 설명할 수 도록 군집의 대표 용어를 추출하는 방법은 다음과 같다. 문서집합을 전처리하여서 용어-문서 빈도행렬 A를 생성하고, 추출하고자 하는 군집의 개수(의미특징 r)를 설정한다. 설정된 군집의 개수를 이용하여서 비음수행렬분해 한다. 행렬분해 된 의미특징행렬 W를 이용하여 군집의 주제를 잘 설명할 수 있는 용어들을 추출한다. 즉, 행렬 W의 열벡터는 군집의 주제에 대응되며, 행벡터는 군집을 구성하는 문서들의 용어에 대응된다. 이러한 이유에서 열벡터에 포함된 높은 값의 의미특징은 그 열벡터에 대응되는 군집에 중요한 용어가 된다. 군집의 대표 용어를 추출하는 식은 다음과 같다.

$$R^p \leftarrow A_{ij} \,\mathrm{if}\, p = \underset{1 \le j \le r}{\arg max}\, W_{ij} \,\mathrm{and}\, W_{ij} \ge cv^j \qquad (5)$$

여기서, R^p는 p번째 군집을 대표하는 용어집합이고, A_{ij}는 j번째 열벡터(군집)에 속하는 i번째 행의 의미특징에 대응되는 용어이다. cv^j는 j번째 열벡터에 포함된 의미특징의 평균값으로 식(6)과 같다.

$$cv^j = \frac{\sum_{i=1}^{n} W_{ij}}{n} \qquad (6)$$

여기서, n은 i행의 개수이다. 즉, 용어(의미특징)의 개수이다. 다음 예1)에서는 본 논문에서 제안한 방법을 이용하여서 군집의 대표 용어를 추출하는 예를 보여준다.

예1) 다음 표 1은 8개의 문서로 구성된 문서집합으로 참고문헌[15]의 그림 4.10의 예제에서 일부 문서를 발취한 예이다. 표 2는 표 1의 문서집합을 전처리하여서 용어-문서 빈도 행렬로 구성한 예이다. 표 3은 표 2의 용어-문서 빈도행렬을 비음수행렬분해 하여서 의미특징

표 1. 7개의 문서로 구성된 문서 집합
Table 1. Document set of composition of 7 documents.

구분	문 서
d1	A course on integral equations
d2	A ttractors for semigroups and evolution equations
d3	Automatic differentiation of algorithms : theory, implementation, and application
d4	Geometrical aspects of partial differential equations
d5	Ideals, varieties, and algorithms - an introduction to computational algebraic geometry and commutative algebra
d6	Oscillation theory for neutral differential equations with delay
d7	Oscillation theory of delay differential equations

표 2. 문서집합(표 1)의 용어-문서 빈도행렬
Table 2. Term-document frequency matrix of document set (Table 1).

문서 / 용어	*d1*	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
course	1	0	0	0	0	0	0
integral	1	0	0	0	0	0	0
equations	1	1	0	1	0	1	1
ttractors	0	1	0	0	0	0	0
semigroups	0	1	0	0	0	0	0
evolution	0	1	0	0	0	0	0
automatic	0	0	1	0	0	0	0
different	0	0	1	1	0	1	0
algorithms	0	0	1	0	1	0	0
theory	0	0	1	0	0	1	1
implementation	0	0	1	0	0	0	0
application	0	0	1	0	0	0	0
geometric	0	0	0	1	1	0	0
aspects	0	0	0	1	0	0	0
partial	0	0	0	1	0	0	0
ideals	0	0	0	0	1	0	0
varieties	0	0	0	0	1	0	0
introudction	0	0	0	0	1	0	0
computational	0	0	0	0	1	0	0
algerbra	0	0	0	0	2	0	0
commutative	0	0	0	0	1	0	0
oscillation	0	0	0	0	0	1	1
neutral	0	0	0	0	0	1	0
delay	0	0	0	0	0	1	1

표 3. 용어-문서 빈도행렬(표 2)을 비음수행렬분해한 의미특징행렬 W와 cv
Table 3. cv and semantic feature matrix W by NMF from term-document frequency matrix (Table 2).

구분	*r1*(군집1)	*r2*(군집2)	*r3*(군집3)
course	0	0.2754	0
integral	0	0.2754	0
equations	0.0183	**2.1334**	0
ttractors	0	0.3162	0
semigroups	0	0.3162	0
evolution	0	0.3162	0
automatic	0.0019	0	**0.9992**
different	0.0628	**0.9209**	**1.0572**
algorithms	**0.9984**	0	**1.0012**
theory	0	**1.0667**	**1.0273**
implementation	0.0019	0	**0.9992**
application	0.0019	0	**0.9992**
geometric	**1.0816**	**0.3434**	0
aspects	0.0851	**0.3816**	0
partial	0.0851	**0.3816**	0
ideals	**0.9965**	0	0.002
varieties	**0.9965**	0	0.002
introudction	**0.9965**	0	0.002
computational	**0.9965**	0	0.002
algerbra	**1.993**	0	0.0039
commutative	**0.9965**	0	0.002
oscillation	0	**1.1603**	0.0281
neutral	0	**0.6329**	0.0774
delay	0	**1.1603**	0.0281
cv	**0.442581**	**0.333157**	**0.296705**

표 4. 표 3으로부터 추출한 군집의 대표 용어 집합
Table 4. The extracted cluster representing term set.

구분	*r1*	*r2*	*r3*
대표용어	algorithms geometric ideals varieties introudction computational algerbra commutative	**equations** different theory geometric aspects partial oscillation neutral delay	automatic different algorithms theory implementation application

행렬 *W*를 생성하고, 생성된 각각의 의미특징 열벡터의 평균값 *cv*를 식(6)을 이용하여 계산한 결과이다. 표 4는 식(5)를 이용하여서 표 3의 의미특징 열벡터에서 열벡터의 평균값보다 큰 의미특징 값에 대응되는 용어를 추출하여서 각각의 군집으로부터 군집의 대표용어를 추출한 결과이다.

나. 워드넷을 이용한 군집의 대표 용어의 확장

워드넷은 영어의 의미어휘목록 도구로 영어 단어를

'synset'이라 불리는 유의어(synonym) 집단으로 분류하여서 간단한 일반적인 정의를 제공한다. 또한 어휘목록 사이의 다양한 의미관계를 포함하고 있으며, 단어집과 시소러스(유의어/반의어)를 구성하여 정보검색 및 인공지능의 다양한 응용분야에 사용할 수 있도록 도구 및 프로그래밍 라이브러리를 지원한다[16].

군집의 대표 용어를 이용하여 문서를 군집할 때, 대표용어와 일치하는 용어들로 구성된 문서들은 잘 군집되나, 대표 용어가 나타내는 군집의 주제를 포함하고 있으면서 다른 용어들로 구성된 문서들은 좋은 군집 결과가 나오지 않는 문제를 가지고 있다. 이러한 문제를 해결하기 위하여 본 논문에서는 워드넷을 이용하여 대표 용어들을 유의어 집단으로 확장한다. 확장방법은 대표 용어를 워드넷을 이용하여서 명사에 대한 유의어를 검색하고, 이 유의어 집합을 대표용어에 추가하여서 확장된 군집주제의 용어 집합 ER^p를 구성한다. 여기서 ER^p는 p번째 군집에서 확장된 군집의 대표 용어 집합 ER이다. 본 논문에서는 명사의 유의어를 검색할 때에 워드넷에서 기본적으로 지원하는 추정 용어 빈도 순위만(the ordered by estimated frequency of noun)을 사용한다. 명사에 대한 추정 용어 빈도 순위만을 사용하는 이유는, 동사나 다른 형태소에 대한 유의어를 대표용어에 추가할 경우, 대표용어가 포함하는 군집의 주제를 너무 많이 벗어나기 때문에 군집의 성능을 오히려 저하 시킨다. 다음 표 5는 표 4의 "equations"를 워드넷 2.1을 이용하여서 명사의 추정 용어 빈도 순위로 나타낸 것이다.

표 5. 표 4의 "equations" 대표용어에 대한 유의어 확장

Table 5. Expanding the representing term of "equations" in Table 4 by synonyms.

유의어
mathematical statement
status, position
equalization, equalisation, leveling

3.3 유사도를 이용한 문서 군집

유사도를 이용한 문서 군집 방법은 다음과 같다. 식 (7)을 이용하여서 각각의 문서와 각각의 군집의 대표 용어집합 R간의 유사도를 계산한다. 군집의 대표 용어 집합과 가장 높은 유사도를 갖는 문서를 대표 용어집합의 군집에 포함시킨다. 그러나 일반 적으로 문서집합에 구성된 문서의 특성들을 보면 군집에 나타내는 주제에 일치하면서 동음이의어(homonym)나 이음동의어(유의어,synonym)으로 구성되어 있어서 유사도를 이용하여 구별할 수 없는 경우가 있다. 이러한 이유에서 본 논문에서는 군집의 대표 용어 집합과의 유사도가 0인 문서가 있다면 확장된 군집의 대표 용어집합 ER과 유사도를 계산하여 군집한다. 다음은 본 논문에서 유사도 계산에 사용되는 코사인 유사도 $csin()$이다[5].

$$csim(A_{*a}, A_{*b}) = \frac{\sum_{i=1}^{m} A_{ia} \times A_{ib}}{\sqrt{\sum_{i=1}^{m} A_{ia}^2} \times \sqrt{\sum_{i=1}^{m} A_{ib}^2}} \quad (7)$$

여기서, A_{*a}와 A_{*b}는 행렬 A의 a번째와 b번째 열벡터이다. 이 것 들은 비음수 값을 가지므로 $0 \leqq csim() \leqq 1$이다.

3.4 제안방법의 알고리즘

다음 그림 4는 본 논문에서 제안한 문서군집방법의 알고리즘이다.

1행에서는 문서집합 D를 전처리하여 용어-문서 빈도 행렬 A를 생성한다. 2행에서는 행렬 A를 비음수행렬분

Algorithm: *SNMF*(D,r);

Input: 문서집합 D, 군집의 개수 r.

Output: 용어(m)-문서(n) 빈도행렬 A, 비음수 행렬 W and H, 군집된 문서집합 C

1: $A \leftarrow$ 전처리(D);
2: $[W, H] \leftarrow$ 비음수행렬분해(A);
3: $cv \leftarrow CV(W)$;
4: for $k \leftarrow 1$ to m do
5: $R^p \leftarrow A_{ij}$ if $p = \underset{1 \le j \le r}{\arg max} W_{ij}$ and $W_{ij} \ge cv^j$;
6: end
7: $ER \leftarrow$ 유의어확장(R);
8: for $k \leftarrow 1$ to n do
9: $C^p \leftarrow A_{*j}$ if $p = \underset{1 \le j \le n, 1 \le k \le r}{\arg max} csin(A_{*j}, R^k)$;
10: if $csin(A_{*j}, R^k) = 0$
11: $C^p \leftarrow A_{*j}$ if $p = \underset{1 \le j \le n, 1 \le k \le r}{\arg max} csin(A_{*j}, ER^k)$;
12: endif
13: end

그림 4. 제안 알고리즘

Fig. 4. The proposed algorithm.

해 하여서 비음수 의미특징행렬 W와 비음수 의미변수 행렬 H를 계산한다. 3행에서는 의미특징 행렬 W와 식(6)을 이용하여서 의미특징 열벡터의 평균 cv를 계산한다. 4행에서 6행까지는 군집의 대표 용어집합을 추출한다. 7행에서는 군집의 대표 용어집합과 워드넷의 유의어 기능을 이용하여서 군집의 대표 용어집합을 확장한다. 8행에서 13행까지는 식(7)의 유사도를 이용하여 문서를 군집한다. 이중 9행에서는 군집의 대표 용어집합과 문서간의 유사도를 이용하여 문서를 군집하며, 10행과 같이 유사도가 0인 경우, 11행에서와 같이 확장된 군집의 대표 용어집합과 문서의 유사도를 이용하여 문서를 군집한다.

Ⅳ. 실험 및 평가

본 논문의 평가자료는 20 Newsgroups문서자료[17]를 이용하였다. 20 Newsgroups문서자료는 문서군집 및 분류의 표준 성능평가 자료로 많이 사용하는 자료이다. 20 Newsgroups는 뉴스 그룹이 20개가 있으며, 20개의 뉴스 그룹에는 총 20000 개의 문서를 포함하고 있다. 뉴스그룹은 컴퓨터 그래픽, 운영체제 윈도우, 컴퓨터 하드웨어, 종교, 의학, 정치 등 20개의 다양한 주제로 구성되어 있으며, 각 주제에 포함된 기사의 수는 같다.

본 논문에서는 문서군집의 성능평가를 위하여 20 Newsgroups문서자료 중 일부를 무작위로 추출하여 사용하였다. 다음 표 6은 평가에 사용된 평가 자료의 특성표이다.

본 논문에서는 성능평가 방법의 척도(measure)로는 식(9)의 NMI(normalize mutual information)를 사용한다[7~9]. NMI는 문서군집의 성능평가에 많이 사용되는 척도로 거의 표준 평가척도 중 하나이다. NMI는 두 군집간의 정보이득을 계산하여서 성능을 평가하는 방법이다. NMI의 상호정보이득은 두 개의 문서군집 C와 C'가 주어질 때 이들 간의 상호정보 MI(C, C')로 다음 식(8)과 같이 정의된다.

표 6. 평가에 사용된 문서집합의 특성
Table 6. The property of document set with respect to evaluation.

문서집합의 속성	20 Newsgroups
총 문서 갯수	20000
사용문서 갯수	5000
클러스터 갯수	20
사용 클러스터 갯수	2~10
최대 클러스터의 문서 갯수	500
최소 클러스터의 문서 갯수	20
중간 클러스터의 문서 갯수	300
평균 클러스터의 문서 갯수	280

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \quad (8)$$

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \quad (9)$$

여기서, $p(c_i)$와 $p(c'_j)$ 는 각각 군집 c_i 와 c'_j 에 문서집합의 문서가 포함될 확률이고, $p(c_i, c'_j)$는 문서집합의 문서가 동시에 군집 c_i 와 c'_j 에 포함될 확률이다. $H(C)$와 $H(C')$는 C와 C'의 엔트로피이다.

본 논문의 실험은 서로 다른 일곱 가지 문서군집방법과 제안방법간의 성능을 비교 평가 하였다. 평가방법은 20 Newsgroups 문자서료로 부터 임의로 추출된 10개의 군집문서를 이용하여서 군집하고, 군집결과를 실제 20 Newgroups에 분류되어 있는 문서와 NMI를 비교하였다. 비교방법으로는 군집의 개수를 2에서 10까지 증가시키며 각각 50번 반복하여서 각각의 군집에 평균을 계산하여서 평가하였다.

평가에 사용된 비교방법들은 직접 구현하였으며, 다음 그림 4와 같이 KM[3~4], NMF[9], ASI[7], CLGR[8], RNMF[10~11], FPCA[13], FNMF[12], SNMF등의 문서군집 방법을 비교 평가 하였다. 여기서 KM은 전통적인 분할 기반의 군집방법으로 Kmeans를 이용한 방법이다. 본 논문에서는 기존 방법과의 비교 기준을 세우기 위하여

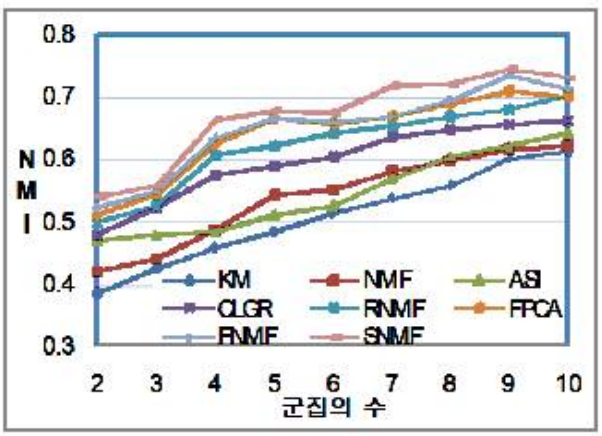

그림 5. 문서군집방법들 간 평균 NMI 비교결과
Fig. 5. The result of comparison of average NMF in document clustering methods.

서 사용되었다. 나머지 NMF, ASI, CLGR, RNMF, FPCA, FNMF, SNMF등은 의미특징을 이용한 방법으로 RNMF, FPCA, FNMF는 이전에 저자들이 제안한 방법이다. 여기서, SNMF는 본 논문에서 제안한 방법이며, FNMF와 FPCA는 비음수행렬분해와 주성분 분석에 각각 퍼지연관을 이용한 문서군집방법들이고, RNMF는 비음수행렬분해와 군집의 정제방법을 이용한 군집방법이다. 또한, NMF는 비음수 행렬분해의 의미특징을 이용한 Xu의 문서군집방법이며, ASI는 Li가 제안한 문서군집방법으로 반복 적응형 군집의 하위 공간 구조를 이용하였으며, CLGR는 Wang이 제안한 방법으로 군집의 지역과 전역의 정규화 속성을 이용하여서 문서를 군집하는 방법이다.

그림 4에서 NMF군집방법이 KM군집방법 보다 성능이 좋은 것은, KM에서의 단순히 두 문서간의 거리 척도를 사용하는 것보다는, NMF의 의미특징들을 이용하여 자료의 내부구조를 반영하는 것이 문서들 간의 관계를 더 잘 표현하여서 군집결과의 정확도에 더 영향을 미치는 것을 알 수 있다[4, 9]. 또한 군집의 하위 공간 구조의 속성을 사용하는 ASI나 군집의 전역 및 지역적 정규화 특성을 사용하는 CLGR보다는 군집의 내부 구조와 군집간의 응집도를 이용하는 RNMF가 좋은 군집결과를 나타냄을 알 수 있다[10~11]. 특히, FPCA나 FNMF는 군집의 각각의 특성을 나태 내는 대표용어와 군집에 포함되는 문서의 용어들 간의 연관관계를 고려함으로써 좋은 성능을 보인다[12~13]. 그러나 FPCA나 FNMF 역시 자료 내부구조의 특성만을 이용하기 때문에 원본 문서자료 집합의 영향을 많이 받는다. 그림 4에서는 제안한 방법인 SNMF가 가장 좋은 성능을 보인다. 이것은 군집의 주제를 잘 표현하는 대표 용어집합과 문서집합의 내부 자료 특성을 고려하여 외부 지식인 워드넷의 유사어를 이용하여서 대표 용어들을 확장함으로써 군집의 특성에 부합되는 문서들을 더 잘 추출하는 것으로 보인다.

V. 결 론

본 논문은 군집 주제의 유의어와 유사도를 이용하여서 문서군집의 결과를 향상시키는 방법을 제안하였다. 제안 방법은 비음수행렬분해를 이용하여서 문서집합의 주제를 잘 표현 할 수 있는 군집 주제의 용어들을 추출하였으며, 비음수행렬의 의미특징이 문서집합의 내부 구조만을 반영하여서 특정 자료 집합에 군집이 제한되는 것을 극복하기 위하여, 워드넷의 유의어를 사용하여 군집 주제의 용어집합을 확장하였다. 또한, 군집 주제 용어와 확장된 용어 집합에 유사도를 이용하여서 문서집합으로부터 군집의 주제를 잘 반영한 문서를 분류하였다. 20 Newsgroups 문서자료를 이용하여서 성능평가한 결과, 제안방법인 SNMF의 평균 NMI가 KM군집 방법에 비하여서는 16.29%가, NMF군집 방법보다는 13.10%가, ASI군집 방법보다는 12.56%가, CRGL군집 방법보다는 7.29%가, RNMF군집 방법보다는 4..73%가, FPCA군집 방법보다는 2.88%가, FNMF군집 방법보다는 2.06%가 각각 높음으로서 다른 문서군집 방법에 비하해서 더 좋은 성능을 나타냄을 알 수 있다. 앞으로 제안 방법의 성능 향상을 위하여 용어에 대한 가중치를 계산할 수 있는 다양한 정책과 다양한 종류의 행렬분해 방법에 적용할 수 있는 방법에 대한 연구가진행 되어야 할 것이다.

참 고 문 헌

[1] S. Basu, A.Banerjee, R. Mooney, “Semi-supervised Clustering by Seeding”, Proceeding of International Conference on Machine Learning (ICML), pp. 19-26, 2002.

[2] X. Hu, X. Zhang, C. Lu, E. K. Park, X. Zhou, “Exploiting Wikipedia as External Knowledge for Document Clustering,” In proceeding of 15th ACM SIGKDD Conference On Knowledge Discover and Data Mining (KDD'09), Paris, Fance, Jun. 2009. pp. 389-396

[3] S. Chakrabarti, “mining the web: Discovering Knowledge from Hypertext Data”, Morgan Kaufmann Publishers, 2003.

[4] J. Han, M. Kamber, “Second Edition Data Mining Concepts and Techniques”, Morgan Kaufman, 2006.

[5] B. Y. Ricardo, R. N. Berthier, “Moden Information Retrieval”, ACM Press, 1999.

[6] D. D. Lee, H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, 401, pp. 788-791, Oct. 1999.

[7] T. Li, S. Ma, M. Ogihara, “Document Clustering via Adaptive Subspace Iteration”, In proceeding of SIGIR'04, pp. 218-225, 2004.

[8] F. Wang, C. Zhang, “Regularized Clustering for Documents”, In proceeding of ACM SIGIR'07, pp. 95-102, 2007.

[9] W. Xu, X. Liu, Y. Gon, "Document Clustering Based On Non-negative Matrix Factorization", Proceeding of Special Interest Group on Information Retrieval (SIGIR), pp. 267-274, 2003.
[10] S. Park, D. U. An, B. R. Char, C. W. Kim, "Document Clustering with Cluster Refinement and Non-negative Matrix Factorization", In proceeding of ICONIP'09, pp. 281-288, 2009.
[11] 박선, 김철원, "비음수 행렬 분해와 군집의 응집도를 이용한 문서군집", 한국해양정보통신학회 논문지, 제13권 제12호, 2603-2608쪽, 2009년.
[12] 박선, 김경준, "비음수 행렬 분해와 퍼지 관계를 이용한 문서군집", 한국항행학회 논문지, 제14권 제2호, 239-246쪽, 2010년.
[13] 박선, 안동언, "주성분 분석과 퍼지 연관을 이용한 문서군집 방법", 한국정보처리학회 논문지, 제17-B권, 제2호, 177-182쪽, 2010년.
[14] 한경한, 남경완, "한국어 정보 처리 입문 : 컴퓨터가 우리말을 이해하려면", 커뮤니케이션북스, 2007년.
[15] W. B. Frankes, B. Y. Ricardo, "Information Retrieval : Data Structure & Algorithms", Prentice-Hall, 1992.
[16] G. Miller, "WordNet: A lexical database for english", CACM, vol. 38(11), 1995, pp.39-41.
[17] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, 2011.

저 자 소 개

박 선(정회원)-교신저자
1996년 전주대학교 전자계산학과 학사 졸업.
2001년 한남대학교 정보산업대학원 정보통신학과 석사 졸업.
2007년 인하대학교 컴퓨터정보공학과 박사 졸업.
2008년～2009년 호남대학교 컴퓨터공학과 전임강사.
2010년 전북대학교 전기전자정보인력양성사업단 박사후과정.
2011년～현재 목포대학교 정보산업연구소 연구교수.
<주관심분야 : 정보검색, 데이터마이닝, 데이터베이스, 해양생물 IT정보융합>

김 경 준(정회원)
1996년 경일대학교 컴퓨터공학과 학사 졸업.
2000년 경북대학교 컴퓨터공학 전공 석사 졸업.
2005년 경북대학교 정보통신학과 박사 졸업.
2005년 경북대학교 컴퓨터공학과 PostDoc.연구원
2005년 대구대학교 정보통신공학부 누리초빙교수
2006년～2009년 호남대학교 전파공학과 전임강사
2009년～현재 한국과학기술원 전산학과 연구교수
<주관심분야 : 센서네트워크, 차세대 인터넷망 구조, 임베디드 소프트웨어>

이 진 석(정회원)
1997년 충남대학교 대학원 무역학과 졸업.
2000년 충남대학교 대학원 무역학과 경영학 석사 졸업.
2007년 충남대학교 대학원 무역학과 경영학박사 졸업.
2002년～현재 정보통신산업진흥원 책임연구원.
<주관심분야 : IT접목서비스, 전자무역, SCM, IT인력양성정책>

이 성 로(정회원)
1987년 고려대학교 전자공학과 졸업
1990년 한국과학기술원 전기및 전자공학과 석사
1996년 한국과학기술원 전기및 전자공학과 박사
1997년 9월～현재 목포대학교 공과대학 정보전자공학과 교수
<주관심분야 : 디지털통신시스템, 이동 및 위성통신시스템, USN/텔레미틱스응용분야, 임베디드시스템>